Comparison of Short-Text Sentiment Analysis Methods for Croatian

Leon Rotim and **Jan Šnajder**
Text Analysis and Knowledge Engineering Lab
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, 10000 Zagreb, Croatia
{leon.rotim,jan.snajder}@fer.hr

Abstract

We focus on the task of supervised sentiment classification of short and informal texts in Croatian, using two simple yet effective methods: word embeddings and string kernels. We investigate whether word embeddings offer any advantage over corpus- and preprocessing-free string kernels, and how these compare to bag-of-words baselines. We conduct a comparison on three different datasets, using different preprocessing methods and kernel functions. Results show that, on two out of three datasets, word embeddings outperform string kernels, which in turn outperform word and n-gram bag-of-words baselines.

1 Introduction

Sentiment analysis (Pang and Lee, 2008) – a task of predicting whether the text expresses a positive, negative, or neutral opinion in general or with respect to an entity – has attracted considerable attention over the last two decades. Some of the more popular applications include political popularity (O'Connor et al., 2010) and stock price prediction (Devitt and Ahmad, 2007). Social media texts, including user reviews (Tang et al., 2009; Pontiki et al., 2014) and microblogs (Nakov et al., 2016; Kouloumpis et al., 2011), are particularly amenable to sentiment analysis, with applications in social studies (O'Connor et al., 2010; Wang et al., 2012) and marketing analyses (He et al., 2013; Yu et al., 2013). At the same time, social media poses a great challenge for sentiment analysis, as such texts are often short, informal, and noisy (Baldwin et al., 2013), and make heavy use of figurative language (Ghosh et al., 2015; Buschmeier et al., 2014).

Sentiment analysis is most often framed as a supervised classification task. Many approaches resort to rich, domain-specific features (Wilson et al., 2009; Abbasi et al., 2008), including surface-form, lexicon-based, and syntactic features. On the other hand, there has been a growing trend in using feature-light methods, including neural word embeddings (Maas et al., 2011; Socher et al., 2013) and kernel-based methods (Culotta and Sorensen, 2004; Lodhi et al., 2002a; Srivastava et al., 2013). In particular, two methods that stand out in terms of both their simplicity and effectiveness are word embeddings (Mikolov et al., 2013a) and string kernels (Lodhi et al., 2002b).

In this paper we focus on sentiment classification of short text in Croatian, a morphologically complex South Slavic language. We compare two simple yet effective methods – word embeddings and string kernels – which are often used in text classification tasks. While both methods are easy to set up, they differ in terms of preprocessing required: word embeddings require a sizable, possibly lemmatized corpus, whereas string kernels require no preprocessing at all. This motivates the main question of our research: do word embeddings offer any advantage over corpus- and preprocessing-free string kernels, and how do these methods compare to simpler bag-of-words methods? To the best of our knowledge, this question has not explicitly been addressed before, especially for a morphologically complex language like Croatian. We present findings from the comparison on three different short-text datasets in Croatian, manually labeled for sentiment polarity, using different levels of morphological preprocessing. To spur further research, we make one dataset publicly available.

2 Related Work

Sentiment classification for short and informal texts has been the focus of considerable research,

Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, pages 69–75,
Valencia, Spain, 4 April 2017.

e.g., (Thelwall et al., 2010; Kiritchenko et al., 2014), especially within the recent SemEval evaluation campaigns (Nakov et al., 2016; Rosenthal et al., 2015; Rosenthal et al., 2014). Recent research has focused on sentence-level sentiment classification using neural networks: Socher et al. (2012) and Socher et al. (2013) report impressive results using a matrix-vector recursive neural network (MV-RNN) and recursive neural tensor networks models over parse trees. Tree kernels present an alternative to neural-based approaches: Kim et al. (2015) and Srivastava et al. (2013) use tree kernels on sentence dependency trees and achieve competitive results. However, as noted by Le and Mikolov (2014), while syntax-based methods work well at the sentence level, it is not straightforward to extend them to fragments spanning multiple sentences. Another downside of these methods is that they rely on parsing, which often fails on informal texts.

Word embeddings (Mikolov et al., 2013a) and string kernels (Lodhi et al., 2002b) present an alternative to syntax-based methods. Tang et al. (2014) and Maas et al. (2011) learn sentiment-specific word embeddings, while Le and Mikolov (2014) reach state-of-the-art performance for both short and long sentiment classification of English texts. Zhang et al. (2008) report impressive performance on Chinese reviews using string kernels.

There has been limited research on sentiment analysis for Croatian. Biđin et al. (2014) applied MV-RNN to prediction of phrase sentiment, while Glavaš et al. (2013) addressed aspect-based sentiment analysis using a feature-rich model. More recently, Mozetič et al. (2016) presented a multilingual study of sentiment-labeled tweets and sentiment classification in different languages, including Croatian. However, they experiment only with classifiers using standard bag-of-words features.

3 Datasets

We conducted our comparison on three short-text datasets in Croatian.[1] The datasets differ in domain, genre, size, and the number of classes. Table 1 summarizes the datasets' statistics.

Game reviews (GR). This dataset originally consisted of longer reviews of computer games, in which annotators have labeled 1858 text spans that express positive or negative sentiment. We used the text spans for our analysis. The spans were labeled by three annotators, and the final annotation was determined by the majority vote on a per-token basis. The spans need not contain full sentences nor need to be limited to a single sentence.

	GR	TD	TG
# Positive	826	2091	2258
# Negative	1032	607	3883
# Neutral	–	269	1858
Total	1858	2967	7999
Avg. # words	7.97	11.12	22.04
Type-token ratio	0.35	0.18	0.21

Table 1: Datasets' statistics

Domain-specific tweets (TD). This dataset contains tweets related to the television singing competition "The Voice of Croatia". The dataset contains 2967 tweets labeled as positive, neutral, or negative by three annotators. The inter-annotator agreement in terms of Fleiss' kappa is 0.721. The final label for each tweet was determined by the majority vote.

General-topic tweets (TG). This is a collection of 7999 general-topic tweets, labeled as positive, neutral, or negative by a single annotator.

The two Twitter datasets, TD and TG, mostly contain informal and often ungrammatical text, whereas the GR dataset is mostly edited, grammatical text. Furthermore, as can be seen from Table 1, Twitter datasets are fairly unbalanced across the three classes, whereas GR is more balanced across the two classes. The GR dataset exhibits the greatest lexical variance, as evidenced by the high type-token ratio. On the other hand, as indicated by the average number of words per text segment/tweet, the texts in TG are longer than the text in the other two datasets.

4 Models

We based all our experiments on the Support Vector Machine (SVM) classification algorithm. Besides being a high-performing algorithm, SVM offers the advantage of using various kernel functions, including string kernels. We used the LIBSVM implementation (Chang and Lin, 2011) for non-linear models and the LIBLINEAR implementation (Fan et al., 2008) for linear models.

Preprocessing. We applied the same preprocessing to all three datasets. For tokenization, we used the Google's SyntaxNet model for Croatian (An-

[1]The Game reviews dataset is available at `http://takelab.fer.hr/croSentCmp`. Due to Twitter terms of use, we do not make other two datasets publicly available.

	GR	TD	TG
# Words	1558	1915	9645
# Lemmas	1383	1484	8101
# Stems	1454	1516	7928
# N-grams	8357	9966	46474

Table 2: BoW baseline feature vector dimensions

dor et al., 2016).[2] Croatian is a highly inflectional language, which has been shown to negatively affect classification accuracy (Malenica et al., 2008). We therefore experimented with two morphological normalization techniques: lemmatization and stemming. For lemmatization, we used the CST lemmatizer for Croatian by Agić et al. (2013). The reported lemmatization accuracy is 97%. For stemming, which is a simple and less accurate alternative to lemmatization, we employed a simple rule-based stemmer by Ljubešić et al. (2007). The stemmer works by stripping the inflectional suffixes of nouns and adjectives. We performed no stopwords removal.

BoW baselines. We evaluated four bag-of-word (BoW) baselines. The baselines use words, stems, and lemmas as features. Additionally, we considered character n-grams, which have been proven useful for text classification of noisy texts (Cavnar et al., 1994). Character n-grams can be viewed as an alternative to morphological normalization, as well as a feature-based counterpart to string kernels. We experimented with 2-, 3-, 4-, and 5-grams, which we combined into a single feature set. From each dataset, we filtered out all words, lemmas, and stems occurring less than two times, and all n-grams occurring less than six times. Table 2 lists the vector feature dimensions after filtering. We used a linear kernel for all baseline models.

Word embeddings. Word embeddings (Mikolov et al., 2013a) belong to a class of predictive distributional semantics models (Turney and Pantel, 2010), which derive dense vector representations of word meanings from corpus co-occurrences. While it has been shown that word embeddings produce high-quality word representations, it has also been shown that they exhibit additive compositionality, i.e., they can be used to represent the compositional meaning of phrases and text fragments by means of simple vector averaging (Mikolov et al., 2013b; Wieting et al., 2015). We trained 300-dimensional skip-gram word embeddings using the word2vec tool[3] on fhrWaC (Šnajder et al., 2013), a filtered version of the Croatian web corpus compiled by Ljubešić and Klubička (2014). We set the window size to 5, negative sampling parameter to 5, and used no hierarchical softmax. When averaging the vectors, we ignored the words, stems, or lemmas that are not covered in the corpus.

SVM's performance very much depends on the choice of the kernel function. For the word embeddings model, we experimented with three different kernels: the linear kernel, the radial basis function (RBF) kernel, and the cosine kernel (Kim et al., 2015). A linear kernel is tantamount to not using any kernel at all and effectively results in a linear model. In contrast, the RBF kernel yields a high-dimensional non-linear model. The cosine kernel is similar to a linear kernel, but additionally includes vector normalization (hence accounting for different-length vectors) and raising to a power:

$$CK(\boldsymbol{x}, \boldsymbol{y}) = \Big[\frac{1}{2}\Big(1 + \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\Big)\Big]^{\alpha}$$

String kernels. A string kernel measures the similarity of two texts in terms of their string similarity, effectively mapping the instances to a high-dimensional feature space. This eliminates the need for features and morphological processing. We experimented with two widely used kernels: a subsequence kernel (SSK) (Lodhi et al., 2002a) and a spectrum kernel (SK) (Leslie et al., 2002). SSK maps each input string s to

$$\varphi_u(s) = \sum_{\boldsymbol{i}: u = s[\boldsymbol{i}]} \lambda^{l(\boldsymbol{i})}$$

where u is a subsequence searched for in s, $\boldsymbol{i}$ is a vector of indices at which u appears in s, l is a function measuring the length of a matched subsequence and $\lambda \leq 1$ is a weighting parameter giving lower weights to longer subsequences. The corresponding kernel is defined as:

$$K_n(s, t) = \sum_{u \in \Sigma^n} \langle \varphi_u(s), \varphi_u(t) \rangle$$

where n is maximum subsequence length for which we are calculating the kernel and Σ^n is a set of all finite strings of length n. The spectrum kernel can be viewed as a special case of SSK where vector of

[2] https://github.com/tensorflow/models/blob/master/syntaxnet/universal.md

[3] https://code.google.com/p/word2vec/

Model/Features	Kernel	GR	TD	TG
BoW baseline				
Words	Linear	0.712	0.673	0.485
N-grams	Linear	0.714	0.690	0.509
Stems	Linear	**0.765**	**0.716**	**0.517**
Lemmas	Linear	0.741	0.711	0.505
Word embeddings				
Words	Linear	0.801	0.653	0.550
Words	RBF	0.807	0.693	**0.565***
Words	Cosine	0.812	**0.715**	0.560
Lemmas	Linear	0.798	0.655	0.536
Lemmas	RBF	0.806	**0.715**	0.543
Lemmas	Cosine	**0.822***	0.711	0.546
String kernels				
–	SK	**0.781**	**0.722**	0.496
–	SSK	0.778	0.718	**0.506**

Table 3: F1-scores for the BoW, word embeddings, and string kernel models on the game reviews (GR), domain-specific (TD), and general-topic (TG) twitter datasets. The best-performing configuration for each model is indicated in bold. Statistically significant differences are marked with *.

indices $\boldsymbol{i}$ must yield contiguous subsequences and λ is set to 1. We compute the string kernels using the Harry string similarity tool.[4]

5 Experiments

Evaluation setup. We evaluated all models using nested k-folded evaluation with hyperparameter grid search (C and γ for RBF, λ and n for SSK, n for SK, α for the cosine kernel). We used 10 folds in the outer and 5 folds in inner (model selection) loop. Following the established practice in evaluating sentiment classifiers (Nakov et al., 2013), we evaluated using the average of the F1-scores for the positive and the negative classes. We used a t-test ($p<0.05$, with Bonferroni correction for multiple comparisons where applicable) for testing the significance of differences between the F1-scores.

Results. Table 3 shows the F1-scores on the three datasets for the baseline, word embeddings, and string kernel models, using different feature sets and kernel configurations. For BoW baselines, the best results are obtained using stemming on all three datasets, i.e., lemmatization does not outperform stemming on neither of the three datasets. For word embeddings, non-linear kernels, cosine kernel in particular, outperform the linear kernel. Lemmatization improves the performance only slightly on the GR dataset, and does not improve or even hurts the performance on the other two datasets. Finally, for string kernels, we obtain the best results with the spectrum kernel on GR and TD datasets, and subsequence kernel on the TG dataset.

Comparing the best results for the three models, we observe that both word embeddings and string kernels outperform the BoW baseline on the GR and TG datasets (statistically significant difference). Overall, word embeddings yield the best performance on these two datasets, while string kernels give the best performance on the TD dataset, though the difference is not statistically significant.

Comparing across the datasets, we notice that the performance on TD and TG datasets is worse than on the GR dataset. This can be traced back to the informality of TD and TG texts, and also the fact that these datasets have three sentiment classes, whereas the GR dataset has only two. The performance on the TG set is probably further impeded by the fact that it covers a variety of topics, and has been annotated by a single annotator.

Discussion. We can make three main observations based on the results obtained. The first is that a word embedding model with a cosine kernel and with either words or lemmas as features significantly outperforms both the baseline and the string kernel model on two out of three datasets. This suggest that a word embedding model should be the model of choice for short-text sentiment analysis in Croatian. The second observation is that lemmatization was mostly not useful in our case: for BoW baseline, stems and n-grams offer better or comparable performance, while for word embeddings lemmatization improved performance on only one out of three datasets. While this could probably be traced back to the noisiness of the informal text (at least for TD and TG datasets), it suggests that lemmatization does not really pay off for this task, especially considering its complexity relative to stemming. Finally, we observe that, although string kernels did not significantly outperform the best baseline models, they do significantly outperform the BoW with words as features on two out of three datasets. Thus, in cases when both a stemmer and word embeddings are not available, string kernels may be the model of choice.

6 Conclusion

We addressed the task of short-text sentiment classification for Croatian using two simple yet effective methods: word embeddings and string kernels.

[4] http://www.mlsec.org/harry/index.html

We trained a number of SVM models, using different preprocessing techniques and kernels, and compared them on three datasets exhibiting different characteristics. We find that word embeddings outperform the baseline bag-of-word-models and string kernels on two out of three datasets. Thus, word embeddings are a method of choice for short-text sentiment classification of Croatian. In cases when word embeddings are not an option, bag-of-words with simple stemming is the preferred method. Finally, if stemming is not available, string kernels should be used. We found lemmatization to be of limited use for this task.

References

Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12.

Željko Agić, Nikola Ljubešić, and Danijela Merkler. 2013. Lemmatization and morphosyntactic tagging of Croatian and Serbian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)*, Sofia, Bulgaria.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how different social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 356–364, Nagoya, Japan.

Siniša Biđin, Jan Šnajder, and Goran Glavaš. 2014. Predicting Croatian phrase sentiment using a deep matrix-vector model. In *Proceedings of the Ninth Language Technologies Conference, Information Society (IS-JT 2014)*, Ljubljana, Slovenija.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49.

William B. Cavnar, John M. Trenkle, et al. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

Ann Devitt and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research*, 9:1871–1874.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. Semeval-2015 Task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478.

Goran Glavaš, Damir Korencic, and Jan Šnajder. 2013. Aspect-oriented opinion mining from user reviews in Croatian. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP)*, pages 18–23, Sofia, Bulgaria.

Wu He, Shenghua Zha, and Ling Li. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3):464–472.

Jonghoon Kim, Francois Rousseau, and Michalis Vazirgiannis. 2015. Convolutional sentence kernel from word embeddings for short text categorization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 775–780, Lisbon, Portugal. Association for Computational Linguistics.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*, pages 538–541, Barcelona, Spain.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of The 31st International Conference on Machine Learning (ICML)*, volume 14, pages 1188–1196.

Christina S. Leslie, Eleazar Eskin, and William Stafford Noble. 2002. The spectrum kernel: A string kernel for svm protein classification. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 7, pages 566–575.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Damir Boras, and Ozren Kubelka. 2007. Retrieving information in Croatian: Building a simple and efficient rule-based stemmer. *Digital information and heritage/Seljan, Sanja*, pages 313–320.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002a. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.

Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. 2002b. Text classification using string kernels. *Journal of Machine Learning Research*, 2(Feb):419–444.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 142–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mislav Malenica, Tomislav Šmuc, Jan Šnajder, and B. Dalbelo Bašić. 2008. Language morphology offset: Text classification on a Croatian–English parallel corpus. *Information processing & management*, 44(1):325–339.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Neural Information Processing Systems Conference (NIPS 2013)*, pages 3111–3119, Lake Tahoe, USA.

Igor Mozetič, Miha Grčar, and Jasmina Smailović. 2016. Multilingual Twitter sentiment classification: The role of human annotators. *PLOS ONE*, 11:1–26.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 Task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 Task 4: Sentiment analysis in Twitter. *Proceedings of SemEval*, pages 1–18.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *International Conference on Web and Social Media (ICWSM)*, pages 122–129, Washington, DC.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. *Proceedings of SemEval*, pages 27–35.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. SemEval-2014 Task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 784–789, Sofia, Bulgaria.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1201–1211. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642.

Shashank Srivastava, Dirk Hovy, and Eduard H. Hovy. 2013. A walk-based semantically enriched tree kernel over distributed word representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1411–1416, Seattle, USA. Association for Computational Linguistics.

Huifeng Tang, Songbo Tan, and Xueqi Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*, 36(7):10760–10773.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1555–1565, Baltimore, MD, USA.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Towards universal paraphrastic sentence embeddings. *arXiv preprint arXiv:1511.08198*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Yang Yu, Wenjing Duan, and Qing Cao. 2013. The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems*, 55(4):919–926.

Changli Zhang, Wanli Zuo, Tao Peng, and Fengling He. 2008. Sentiment classification for Chinese reviews using machine learning methods based on string kernel. In *Proceedings of the 3rd International Conference on Convergence Information (ICCIT)*, volume 2, pages 909–914, Busan, Korea. IEEE.